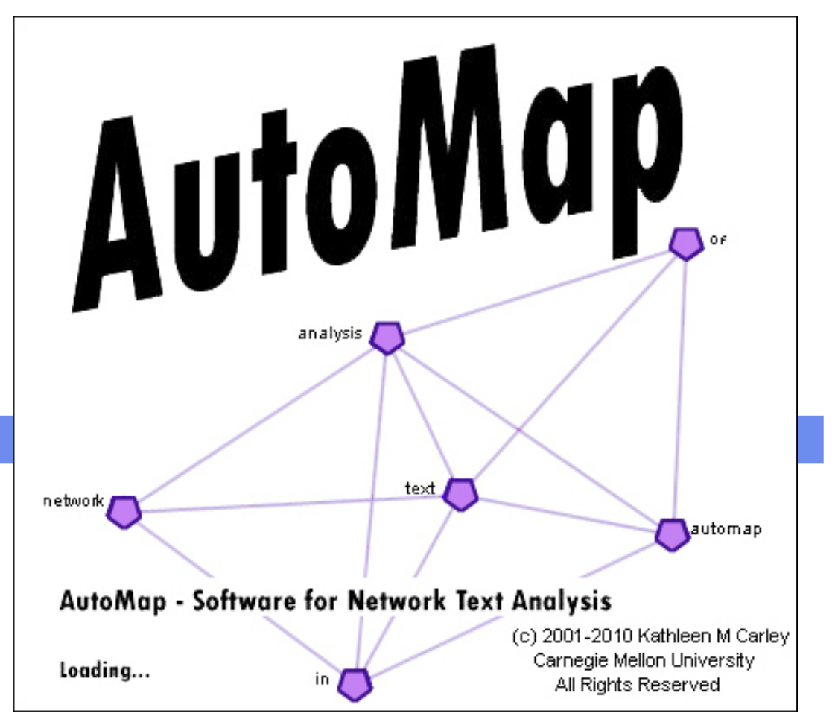
CASOS Summer Institute 2010

AutoMap:

Extracting usable information from unstructured texts


AutoMap
of
analysis
network
text
automap
in
AutoMap - Software for Network Text Analysis
Loading...
(c) 2001-2010 Kathleen M Carley
Carnegie Mellon University
All Rights Reserved

Michael W. Bigrigg
bigrigg@cs.cmu.edu

Prof. Kathleen M. Carley
kathleen.carley@cs.cmu.edu

Frank Kunkel
fkunkel@cs.cmu.edu

Jana Diesner
diesner@cs.cmu.edu

Todd Eisenberg
teisenbe@andrew.cmu.edu

Daniel Chieffallo
dlc85@cs.cmu.edu

Dave Columbus
columbus@cs.cmu.edu

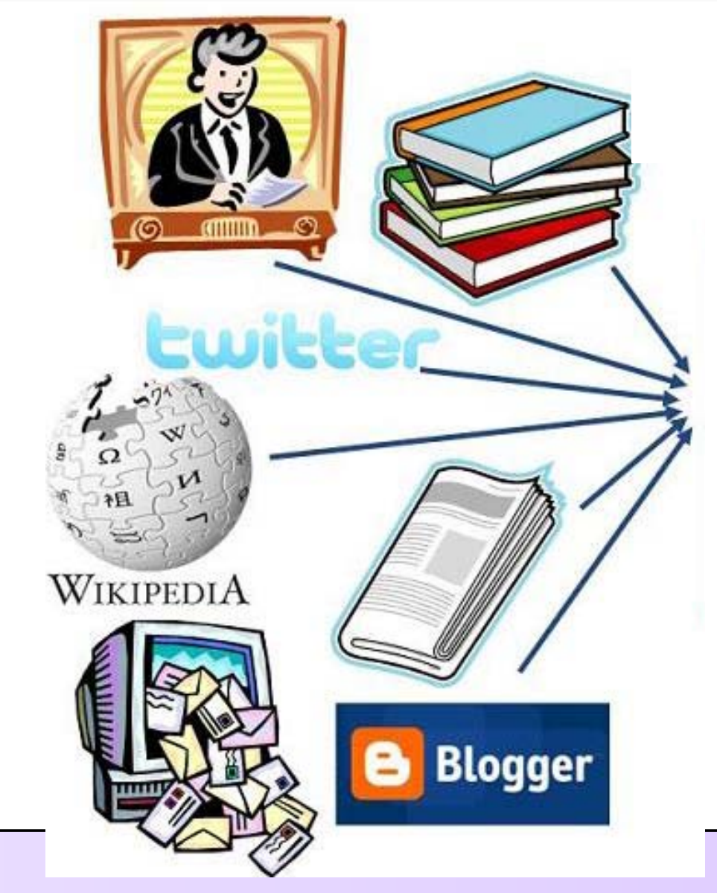
AutoMap is a text mining tool that enables the extraction of network data from texts. AutoMap can extract four types of information:

- content (concepts, frequencies and meta-data such as sentence length)
- semantic networks (concepts and relationships)
- meta-networks (ontologically coded concepts and relationships – named entities and links)
- sentiment and node attributes (attributes of named entities)

Parts of speech tagging, machine learning techniques and proximity analysis enable computer-assisted Network Text Analysis (NTA). NTA encodes the links among words in a text and constructs a network of the linked words.

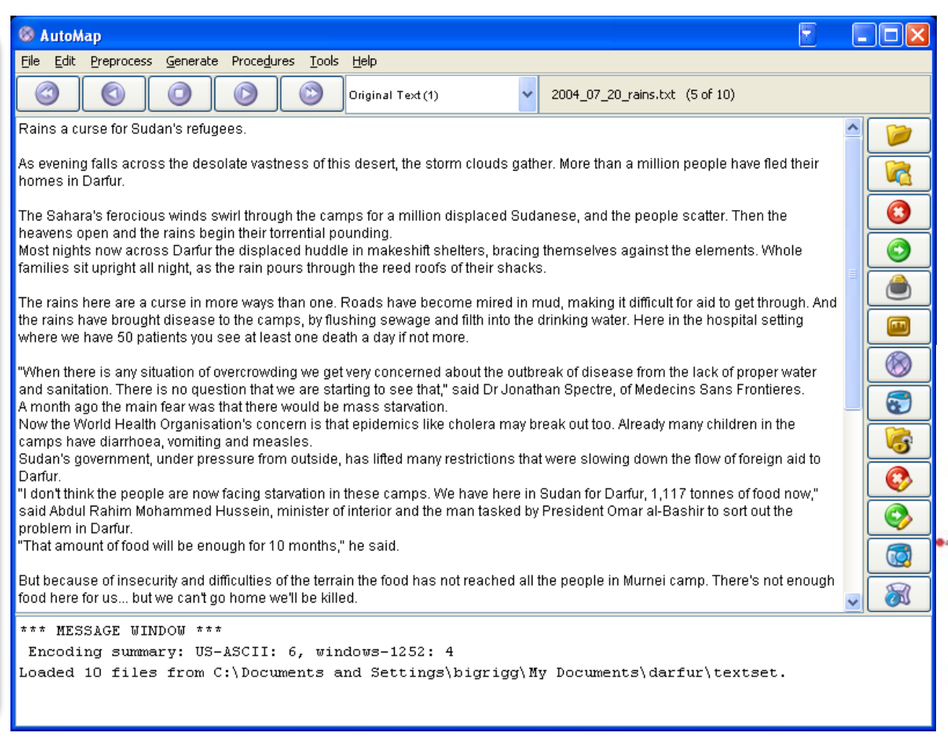
AutoMap is available as a stand-alone application as well as a module in *ORA & SORASCS.

Unstructured Data: Texts



Preprocessing

AutoMap: Data Extraction


AutoMap
File Edit Preprocess Generate Procedures Tools Help
Original Text (1)
2004_07_20_rains.txt (5 of 10)

Rains a curse for Sudan's refugees.

As evening falls across the desolate vastness of this desert, the storm clouds gather. More than a million people have fled their homes in Darfur.

The Sahara's ferocious winds swirl through the camps for a million displaced Sudanese, and the people scatter. Then the heavens open and the rains begin their torrential pounding.
Most nights now across Darfur the displaced huddle in makeshift shelters, bracing themselves against the elements. Whole families sit upright all night, as the rain pours through the reed roofs of their shacks.

The rains here are a curse in more ways than one. Roads have become mired in mud, making it difficult for aid to get through. And the rains have brought disease to the camps, by flushing sewage and filth into the drinking water. Here in the hospital setting where we have 50 patients you see at least one death a day if not more.

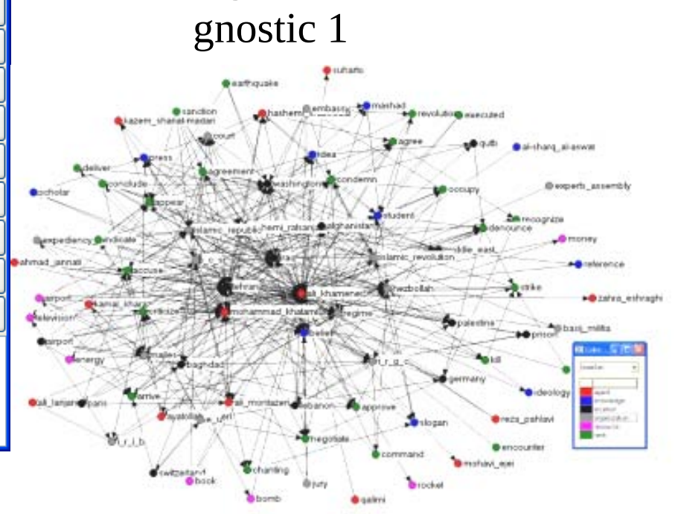
"When there is any situation of overcrowding we get very concerned about the outbreak of disease from the lack of proper water and sanitation. There is no question that we are starting to see that," said Dr Jonathan Spectre, of Medecins Sans Frontieres.
A month ago the main fear was that there would be mass starvation.
Now the World Health Organisation's concern is that epidemics like cholera may break out too. Already many children in the camps have diarrhoea, vomiting and measles.
Sudan's government, under pressure from outside, has lifted many restrictions that were slowing down the flow of foreign aid to Darfur.
"I don't think the people are now facing starvation in these camps. We have here in Sudan for Darfur, 1,117 tonnes of food now," said Abdul Rahim Mohammed Hussein, minister of interior and the man tasked by President Omar al-Bashir to sort out the problem in Darfur.
"That amount of food will be enough for 10 months," he said.

But because of insecurity and difficulties of the terrain the food has not reached all the people in Murnei camp. There's not enough food here for us... but we can't go home we'll be killed.

```
*** MESSAGE WINDOW ***
 Encoding summary: US-ASCII: 6, windows-1252: 4
Loaded 10 files from C:\Documents and Settings\bigrigg\My Documents\darfur\textset.
```

CSV & DyNetML Output

the 20
a 30
John_Doe 15
cattle 10
bars 8
argument 1
gnostic 1



Post Processing

- •Event inference
- •Belief inference
- •Node attribute data fusion
- •Lat-Lon data fusion

ORA: Analysis

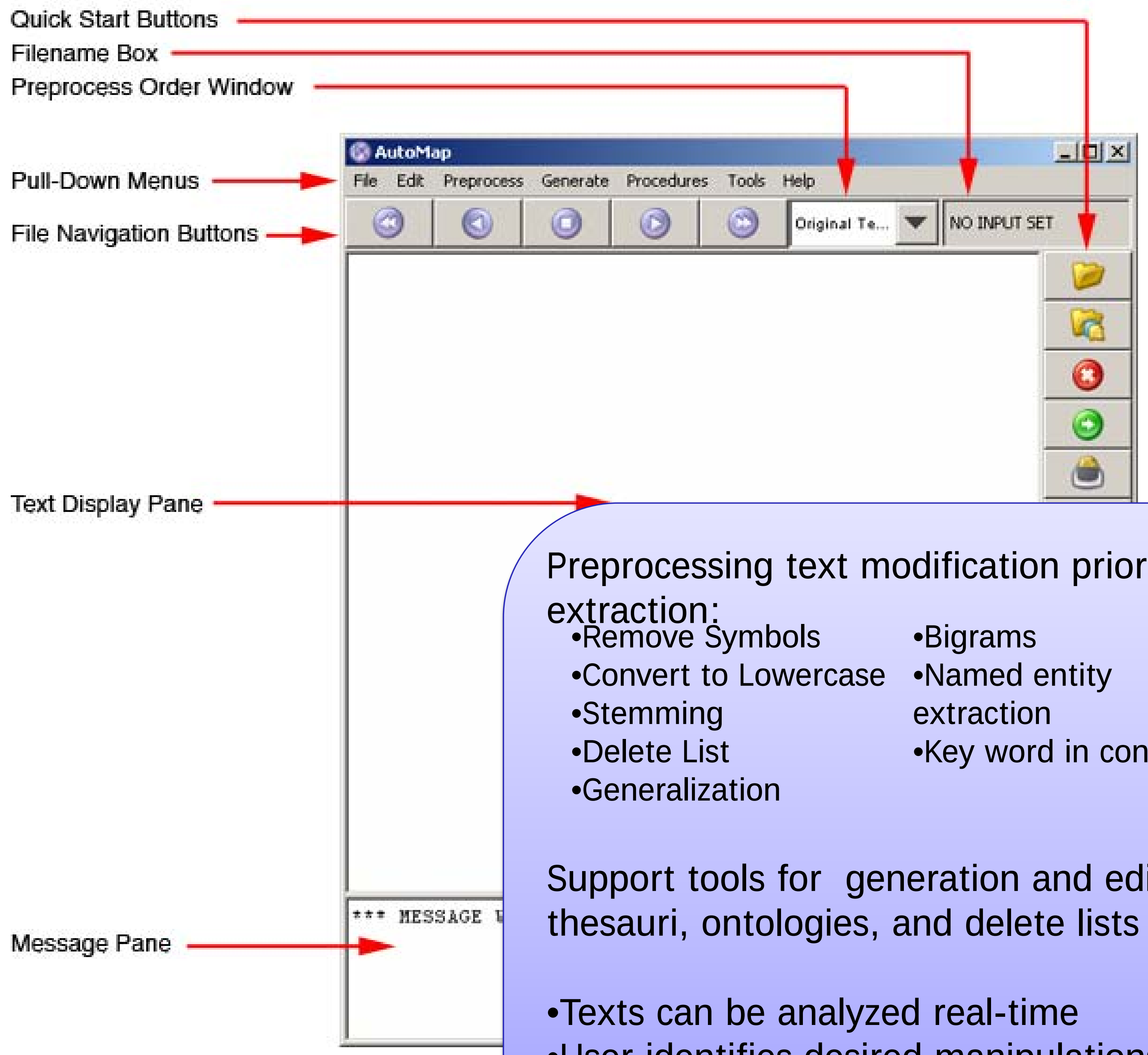
Specialized Text analytics:

- Theme identification – LSA
- Hot Topics
- Semantic Network
- Parts of Speech
- Communicators
- Communicative Power

And – all social and dynamic network tools

Extraction of the network is done using AutoMap. Additional Information added with Post processing Data fusion Routines. Analysis and visualization of networks is done using *ORA.

The GUI Interface



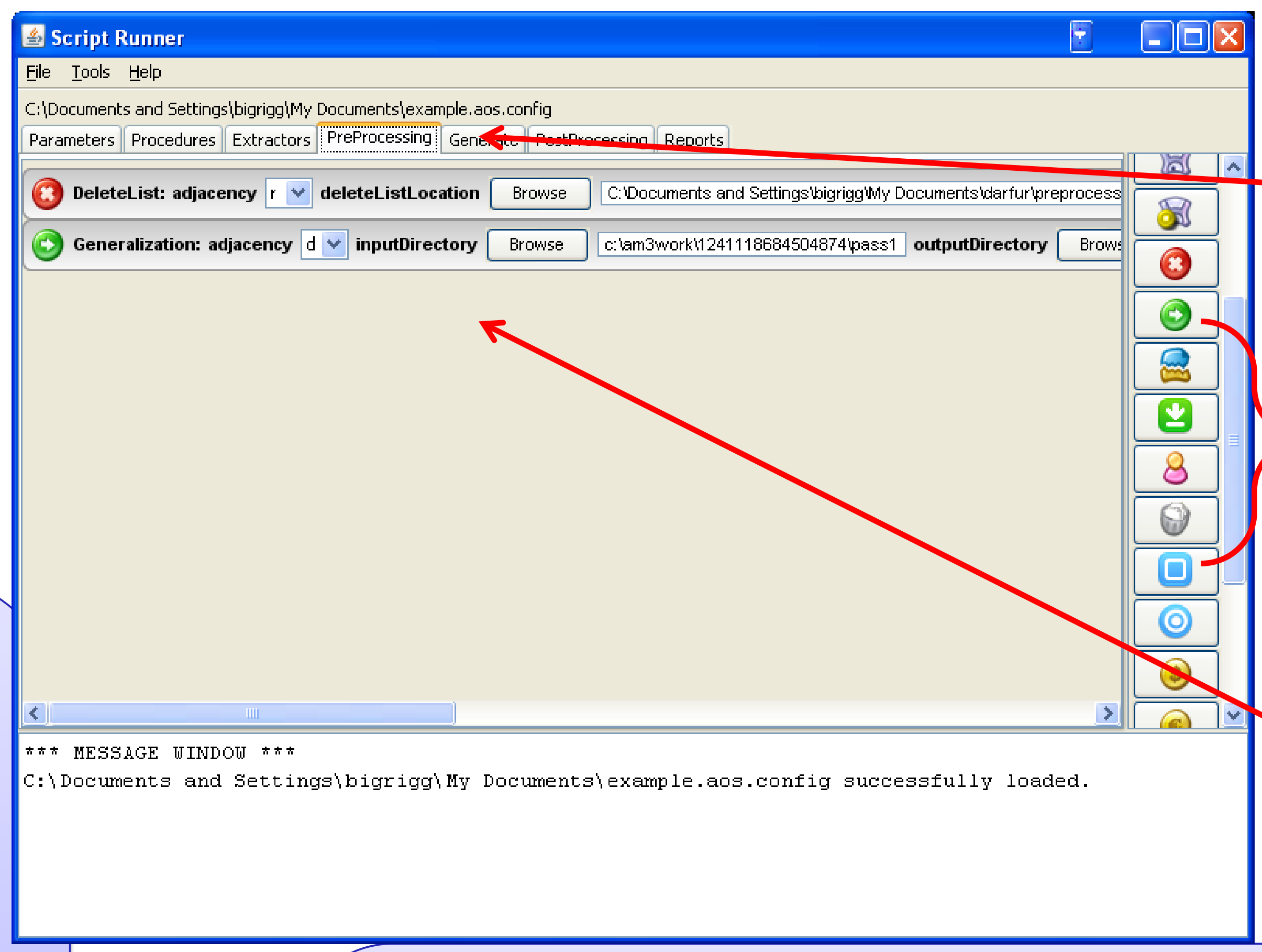
Preprocessing text modification prior to data extraction:

- •Remove Symbols
- •Convert to Lowercase
- •Stemming
- •Delete List
- •Generalization
- •Bigrams
- •Named entity extraction
- •Key word in context

Support tools for generation and editing of thesauri, ontologies, and delete lists exist

- •Texts can be analyzed real-time
- •User identifies desired manipulations in the GUI
- •Manipulations can be saves as a workflow script
- •This script can be run in batch mode

ScriptRunner Interface
Draq & drop editor to adjust workflow



C:\Documents and Settings\bigrigg\My Documents\example.aos.config

DeleteList: adjacency r deleteListLocation C:\Documents and Settings\bigrigg\My Documents\darfur\preprocess

Generalization: adjacency d inputDirectory c:\am3work\1241118684504874\pass1 outputDirectory

```
*** MESSAGE WINDOW ***
C:\Documents and Settings\bigrigg\My Documents\example.aos.config successfully loaded.
```

Multiple tabs for different phases of processing

Palette of available commands.

Canvas of the current commands. Drag and drop new items here.

- •The ScriptRunner can read AutoMap scripts
- •Using Scriptrunner the user can:
  - •Add in webscrapers
  - •Add post-processing options e.g. add lat-lon from gazateers
- •The SuperScript module utilizes all processors in a multi-core environment, paralellizing the text processing

This work is part of the Dynamics Networks project at the center for Computational Analysis of Social and Organizational Systems (CASOS) of the School of Computer Science (SCS) at Carnegie Mellon University (CMU). Support was provided, in part, by the Office of Naval Research (ONR HSCB N000140811223, MMV - N00014-06-1-0104, MURI N000140811186) , the Army Research Institute (W91WAW07C0063) and CASOS. The views and conclusions contained in this document are those of the author sand should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research, the Army Research Institute or the U.S. government.

www.casos.cs.cmu.edu